Serving national scientific communities - genome analysis as an example

Craig A. Stewart, Ph.D.

Orcid ID: 0000-0003-2423-9019

Executive Director, Pervasive Technology Institute

Associate Dean, Research Technologies

Indiana University

Please cite as: Stewart, C.A. 2014. Serving national scientific communities - genome analysis as an example. ZKI-Frühjahrstagung, Berlin, Deutschland. 25 March 2014. http://http://hdl.handle.net/2022/17383

License specifics in last slide

A bit about myself

Biologist by training

Information technologist by luck, and change in NSF funding priorities

Married to a lovely German lady almost 30 years

Began coming to Germany for supercomputing conferences in 1997 and have continued visiting and collaborating ever since

Former Chairperson of the Coalition for Academic Scientific Computation (a group much like this one)

Now lead the Pervasive Technologies Institute and the Research Technologies Division of University Information Technology Services at Indiana University

Principal Investigator for National Center for Genome Analysis Support; NCGAS funds a graduate student who works with iPlant; IU subcontract PI for XSEDE

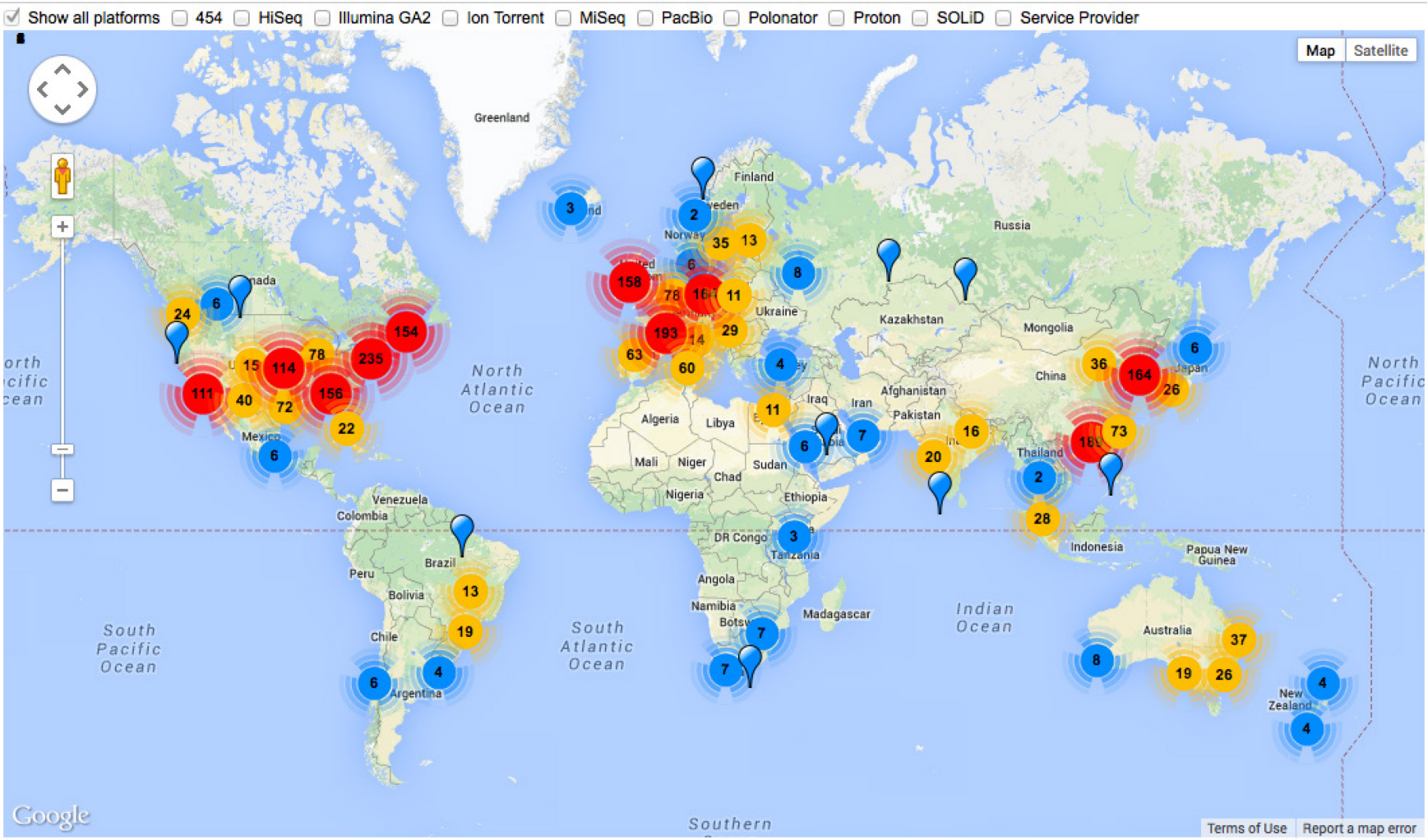
Outline

- Drivers of our current needs – next-generation sequencers
- A tale of three projects (XSEDE, iPlant, National Center for Genome Analysis) – service and service integration for the national community
- Thoughts about how we support science with computing in the future

In this talk, I contend that:

- Serving national communities is best done by multiple, interacting projects
- Software innovation and software support are two different processes
- While the political and financial models for supporting research in the US and EU are very different, there are – I hope – lessons that can be applied to supporting research communities in Germany and in the EU based on experiences in the US

Next-generation Sequencers Worldwide



02/26/2014 http://omicsmaps.com/ (no copyright terms specified)

Dealing with language

Cyberinfrastructure (primarily a US term): “Cyberinfrastructure consists of computing systems, data storage systems, advanced instruments and data repositories, visualization environments, and people, all linked together by software and high performance networks to improve research productivity and enable breakthroughs not otherwise possible.” (Stewart, 2007)

eScience (primarily an EU term): “In the future, e-Science will refer to the large scale science that will increasingly be carried out through distributed global collaborations enabled by the Internet. Typically, a feature of such collaborative scientific enterprises is that they will require access to very large data collections, very large scale computing resources and high performance visualization back to the individual user scientists.” (National e-Science Centre, 2010)

Some history

Decade	Bioinformatics & genome analysis	Cyberinfrastructure
1980s	FASTA, BLAST	NSF supercomputer center program ("time machines"): 1985 - 1997
1990s	Expansion of bioinformatics software; genome assemblers	Partnerships for Advanced Computational Cyberinfrastructure – 1997-2004
2000s	2001 – first publication of human genome More species sequenced; progressive advances in sequencing technology **iPlant initially funded in 2008;** **Next generation sequencers: faster, shorter reads, with errors**	Grid computing is key meme for decade LeMieux @ PSC in 2000 TeraGrid – 2001 ETF in 2001 Terascale Extensions 2003; TeraGrid in production 2004 2003 Atkins report, thins strategic plan for Cyberinfrastructure
2010s	Sequencing becomes relatively inexpensive; **DNA assemblers require large amounts of memory.** NCGAS funded in 2011	Cloud computing to be key meme for decade? ***XSEDE succeeds TeraGrid – supports productivity as well as capability***

Existing government dictates and realities

Dictates:

- Open Data – eventually
- Grant awards are made to universities and colleges

Realities:

- Some data collected now will be of value indefinitely.
- No one has a good solution to long-term storage of data, but funding agencies view it as a problem owned by the universities.
- Federal funding agency budgets are not sufficient to solve the problems of persistent storage of data and data openness.
- Many experiments are going on relative to data curation.
- There is considerable public skepticism about publically funded research in the US. Some is not deserved, some may be, but more pressure is a reality.
- No one is really sure what to do about so-called “cloud computing.”

National infrastructure serving genome science

Creators of new software

NCGAS – small, serving large community largely reactively

- Trinity
- Galaxy
- ABySS
- Velvet

iPlant – large collaborative serving plant science

- DNA Subway
- iPlant Discovery Environment
- Many bioinformatics software applications planned as part of group strategy

XSEDE – designed to serve all research communities

- Stampede
- Gordon
- Blacklight
- Comet
- Mason
- Wrangler
- FutureGrid

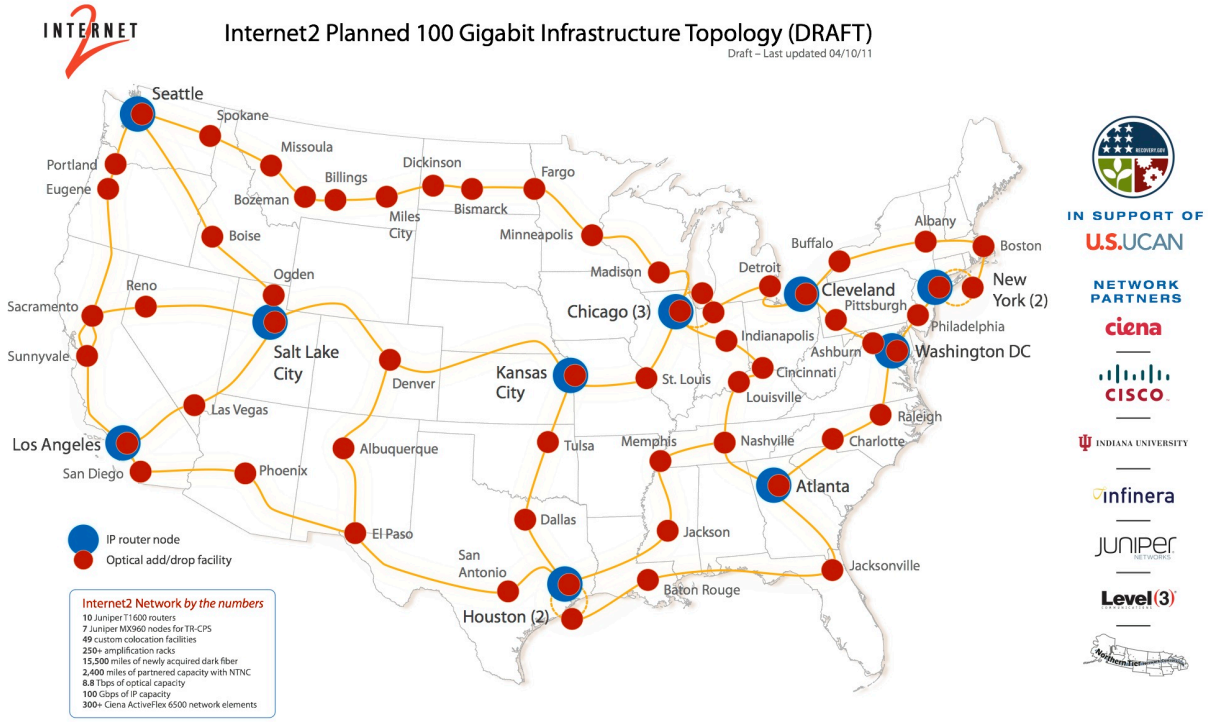
Network – essentially independent of any particular research community

- Internet2
- Regional providers

Internet2

Created as a consortium of universities and colleges in the US to provide an affordable national network.

Internet2 is now very involved in value-added services (such as identity management through InCommon).



XSEDE (eXtreme Science and Engineering Discovery Environment)

XSEDE

XSEDE (eXtreme Science and Engineering Discovery Environment)

XSEDE's Mission: To substantially enhance the productivity of a growing community of researchers, engineers, and scholars through access to advanced digital services that support open research; and to coordinate and add significant value to the leading cyberinfrastructure resources funded by the NSF and other agencies.

Strategic Goals: XSEDE will deepen the use of the advanced digital research services ecosystem … prepare the current and next generation of researchers, engineers, and scholars in the use of advanced digital technologies via education, training, and outreach; and we will raise the general awareness of the value. … XSEDE will sustain the advanced digital research services ecosystem…

Budget: ~$25M US/year ($123M US over 5 years for coordination and support of resources)

Addition of new resources varies with National Science Foundation budgets and balance between hardware and software.

XSEDE resources

System	Type of resource	Type of Service Provider?
Stampede	Large-scale distributed memory parallel	NSF funded, Level 1
Gordon	Large-scale distributed memory parallel, pseudo-large memory	"
Blacklight	Large memory	"
Comet	New - VMs	"
Wrangler	Storage	"
FutureGrid	Experimental computer science / cloud system	"
Mason	*Large memory (low cores)*	*IU-funded, Level 2*
Rockhopper	*Commercial "cluster as a service" owned by Penguin Computing and housed at / supported by IU*	*Commercially owned, Level 3*

Pittsburgh Supercomputing Center Blacklight (SGI Altix® UV 1000) - Massive Coherent Shared Memory Computer

- **2×16 TB of cache-coherent shared memory, 4096 cores**
  - *ideal for genome sequence assembly*
  - High bandwidth, low latency interprocessor communication
- **SUSE Linux operating system**
  - *excellent for portability:* supports OpenMP, C, C++, Java, Perl, Python, p-threads, MPI, UPC
  - *rapid algorithm development*

This slide courtesy Philip Blood, Pittsburgh Supercomputing Center, © PSC

Mason (hp) @ IU

Supports data-intensive high performance computing tasks for IU researchers, faculty, staff, and students on all campuses.

Specs:

- Peak performance of 3.83 teraFLOPS
- 8 TB total RAM - 512 GB RAM per node – really a system of memory with a few processors attached
- Uses Lustre/Data Capacitor II as high-performance file system
- Connects to IU's high-speed research network via 10 Gbps connection

iPlant

Following slides courtesy Steve Goff, PI, iPlant

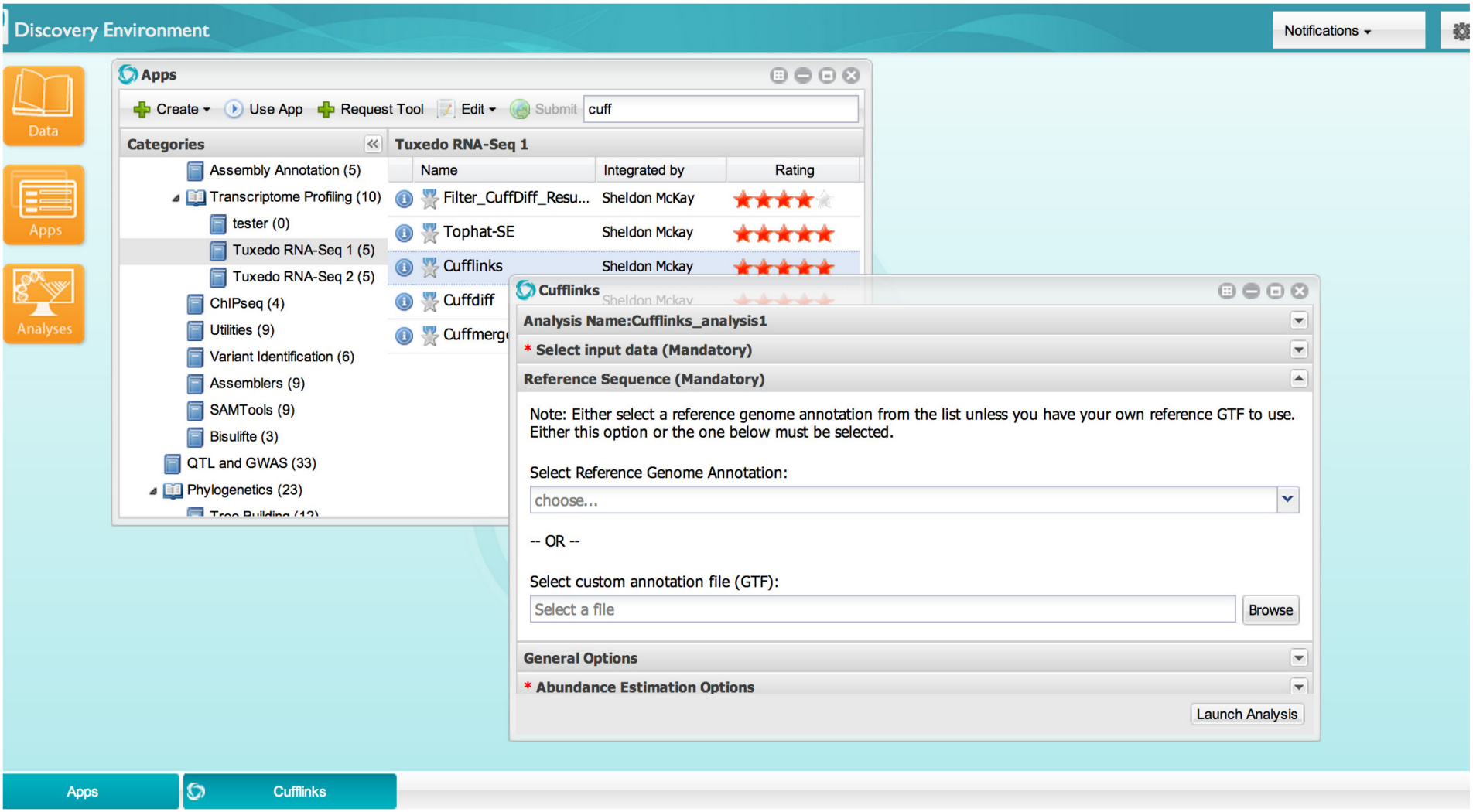
(March 2014)

iPlant – Plant Cyberinfrastructure

Goals:

- *"to create a new type of organization — a cyberinfrastructure collaborative for plant science"*
- *"to enable new conceptual advances through integrative, computational thinking"*
- *"to address an evolving array of grand challenge questions in plant science: the driving force and organizing principles for the collaborative"*
- ~ $10M / year ($50M NSF-funded Project – 5 years, renewed in 2013)
- iPlant is a cyberinfrastructure *platform*
- The *platform* is developed by iPlant and extensible by users
- NSF recommended scope beyond plants.

iPlant Discovery Environment – > 400 applications
VMs promote replicability (RNA-Seq as example)



iPlant Collaborative™

iPlant Cloud Services PROJECT ATMOSPHERE

Customized cloud platform for computing on your terms !

New biology priorities going forward:

- Expand Scope to Non-plant Species
- Continue Support for NGS
- Deliver CI Platform for Modeling, Molecular Breeding
- Expand Support for Ecophysiology
- Continue Range Map Creation for Biodiversity
- Integrate Environmental Information
- Support Additional Molecular Profiling Tech

XSEDE Novel and Innovative Projects program

- Novel and Innovative Projects within XSEDE is intended to be reactive to new user needs, with current focus on life sciences
- Works with developers to port key de novo assembly applications to large, shared-memory system, Blacklight
- Availability of Blacklight highlighted on Broad Institute developer web pages (ALLPATHS-LG and Trinity) and genomeweb.com
- Enthusiastic response from research community – dozens of new groups using Blacklight for de novo assembly every year
- Example projects:
  - **Cold Spring Harbor:** Assembled **5 and 10 gigabase wheat species** using **3 and 6 TB RAM** respectively. Targeting assembly of **16 gigabase** wheat genome (**ALLPATHS-LG**).
  - **Cornell** and **Broad Institute:** Assembled **20 primate transcriptomes at ~1 TB RAM each** (**Trinity**). Understanding evolutionary processes and gaining insight into human disease.

This slide courtesy Philip Blood, Pittsburgh Supercomputing Center, © PSC

NCGAS – National Center for Genome Analysis Support

- Mason provided as “facilities” with IU funding, for use by national research community, through XSEDE, as part of this award
- IU also hosts the commercially owned Rockhopper system – owned and managed by Penguin Computing, a “pay to use” system, software installed and supported by NCGAS
- ~ $0.7M / year budget (award of $1.5M over 3 years + match)
- Focused on user-driven needs
- ~ 4 FTEs (Full Time Equivalents = 1 person) total
- Newest of the projects discussed – funded starting in 2011 (implies that situation prior to 2011 was not optimal)

National Center for Genome Analysis Support
"Mind the Gap"

Gap	How we fill it
System configurations offered by XSEDE and what people doing genome assembly need	Mason (IU contribution to facilities)
Software on XSEDE is not what people need	NCGAS installs and maintains
Software works slowly	NCGAS tunes / re-engineers
People just need help	NCGAS provides consulting NCGAS goes to conferences and informs people about our services
People need storage	***NCGAS provides tape storage (IU facilities)***
People need to publish data sets	***IU provides resources via IUScholarWorks***

NCGAS role in research in general

Bioinformatics should be available to any researcher who is knowledgeable about the biology, regardless of their background in informatics or computer related fields.

For those who know where and how they want to accomplish their analyses, we step back and let them do their science.

For those who need advice, a place to start, have never used a Unix shell, or are not sure whether this parameter or that will provide a better result – we are standing by to help.

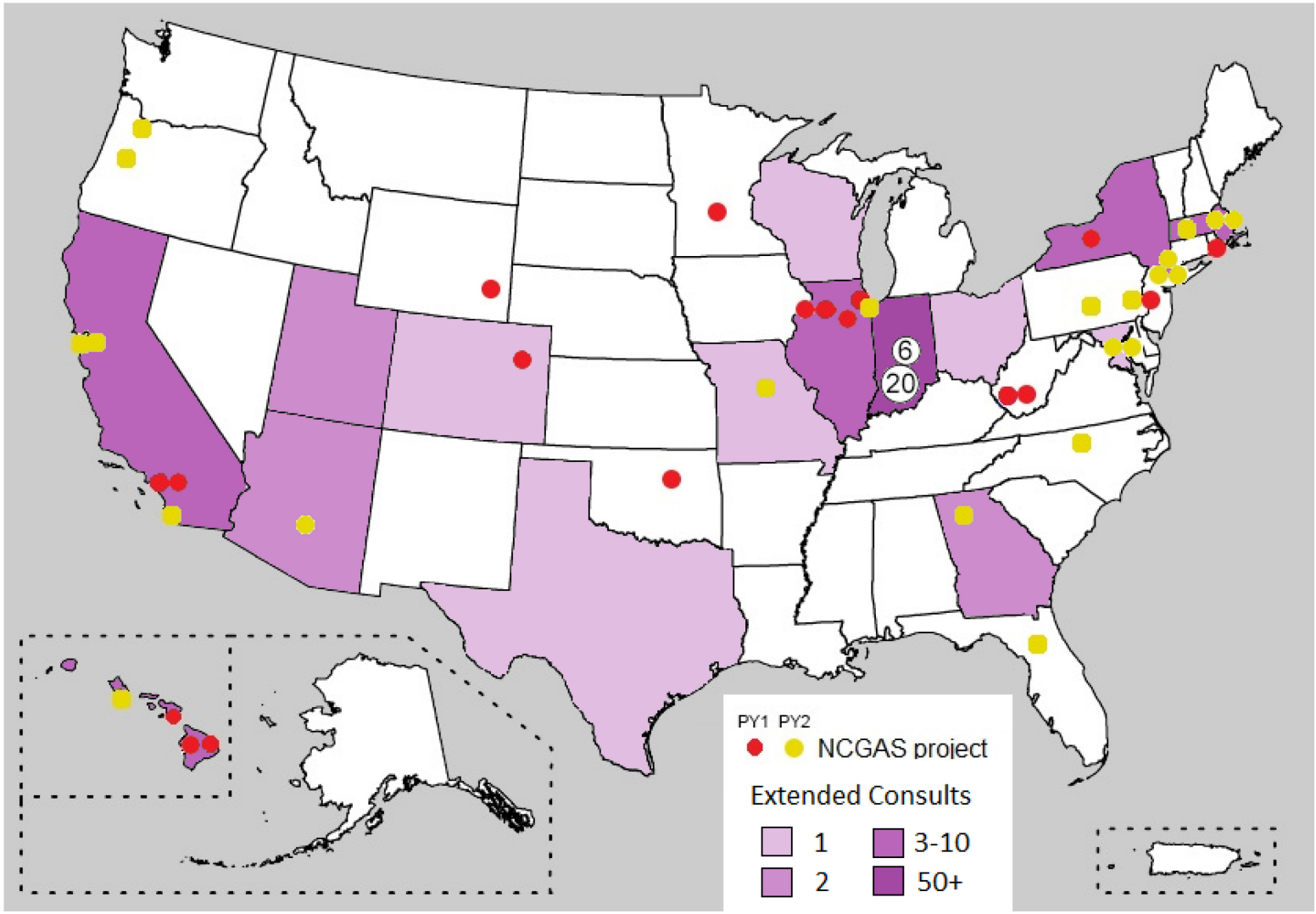
What is the difference in user friendliness?

$

vs.

>

Projects and extended consultations



NCGAS assistance - Dr. Melissa Pespini's research on dung beetles

We recommended assembly procedures and Unix commands – when and how to concatenate data sets together to retrieve desired information.

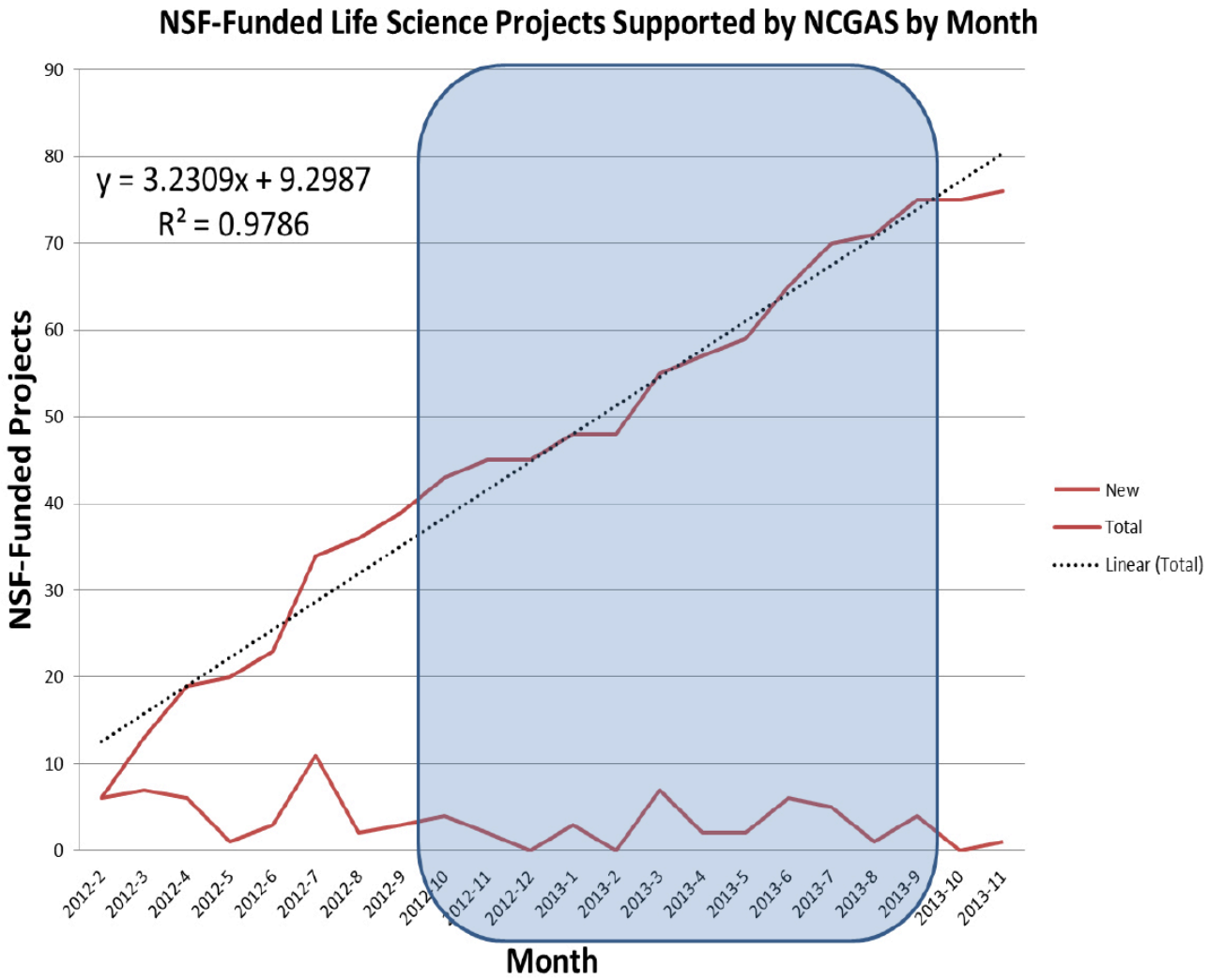
We solved issues with the system that were beyond user experience.

We added new users and brought them up to speed on the project and on Unix.

We wrote customized scripts to get the data in the format the requested programs required.

We assisted with the data-moving process and advised steps to take upon data corruption and failures.

en.wikipedia.org/wiki/File:Scarabaeus_viettei_01.jpg licensed under the Creative Commons Attribution-Share Alike 3.0 Unported, 2.5 Generic, 2.0 Generic and 1.0 Generic license.


NSF-Funded Life Science Projects Supported by NCGAS by Month
y = 3.2309x + 9.2987
R² = 0.9786
NSF-Funded Projects
Month
New
Total
Linear (Total)
0
10
20
30
40
50
60
70
80
90
2012-2
2012-3
2012-4
2012-5
2012-6
2012-7
2012-8
2012-9
2012-10
2012-11
2012-12
2013-1
2013-2
2013-3
2013-4
2013-5
2013-6
2013-7
2013-8
2013-9
2013-10
2013-11

Trinity – not even part of the discussion in proposal for NCGAS

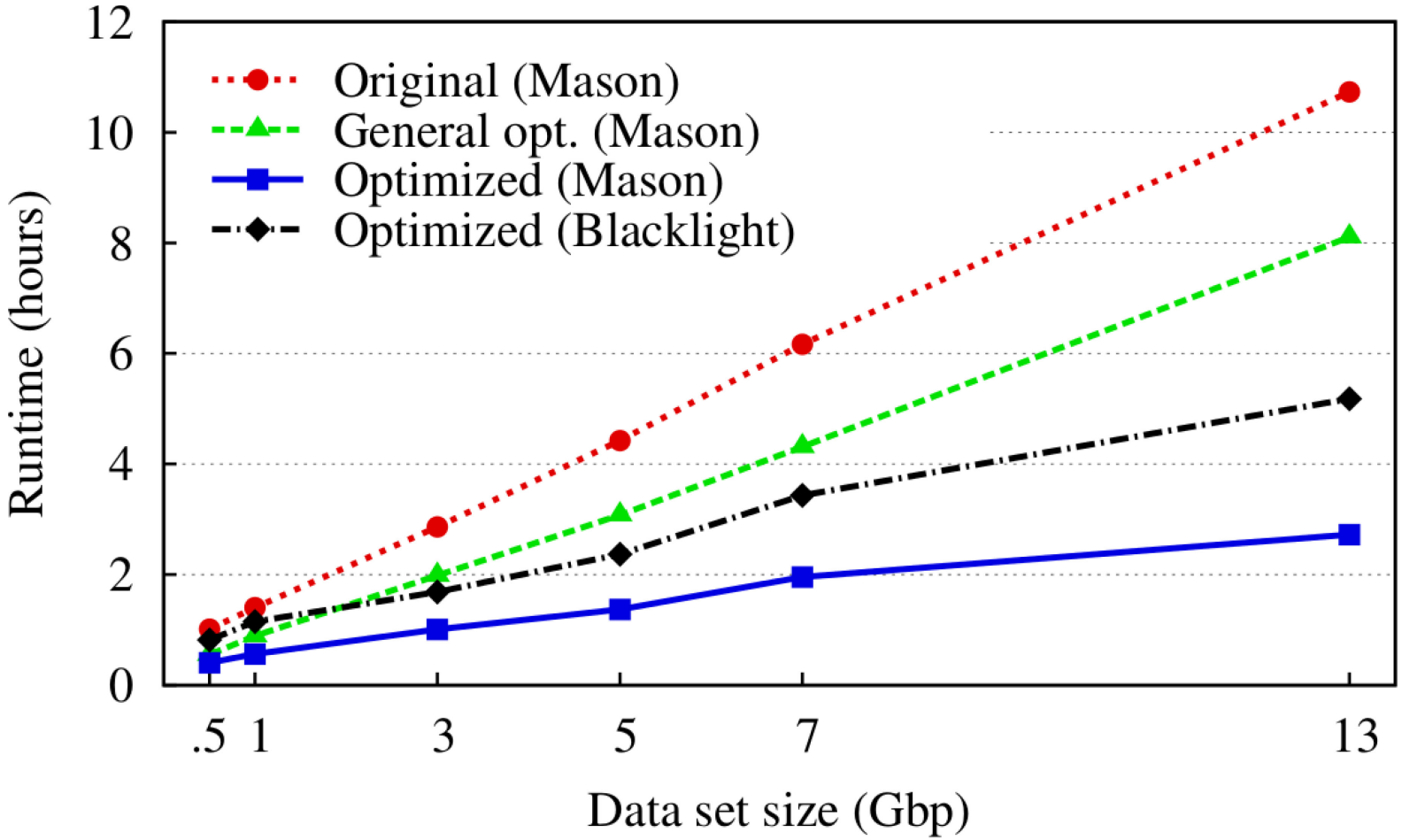
RNA-Seq De novo Assembly Using Trinity

Trinity, developed at the Broad Institute and the Hebrew University of Jerusalem, represents a novel method for the efficient and robust de novo reconstruction of transcriptomes from RNA-seq data. Trinity combines three independent software modules: Inchworm, Chrysalis, and Butterfly, applied sequentially to process large volumes of RNA-seq reads. Trinity partitions the sequence data into many individual de Bruijn graphs, each representing the transcriptional complexity at at a given gene or locus, and then processes each graph independently to extract full-length splicing isoforms and to tease apart transcripts derived from paralogous genes. Briefly, the process works like so:

- **Inchworm** assembles the RNA-seq data into the unique sequences of transcripts, often generating full-length transcripts for a dominant isoform, but then reports just the unique portions of alternatively spliced transcripts.
- **Chrysalis** clusters the Inchworm contigs into clusters and constructs complete de Bruijn graphs for each cluster. Each cluster represents the full transcriptonal complexity for a given gene (or sets of genes that share sequences in common). Chrysalis then partitions the full read set among these disjoint graphs.
- **Butterfly** then processes the individual graphs in parallel, tracing the paths that reads and pairs of reads take within the graph, ultimately reporting full-length transcripts for alternatively spliced isoforms, and teasing apart transcripts that corresponds to paralogous genes.

From: http://trinityrnaseq.sourceforge.net - no copyright terms stated

Final Results – code contributed to definitive Trinity release



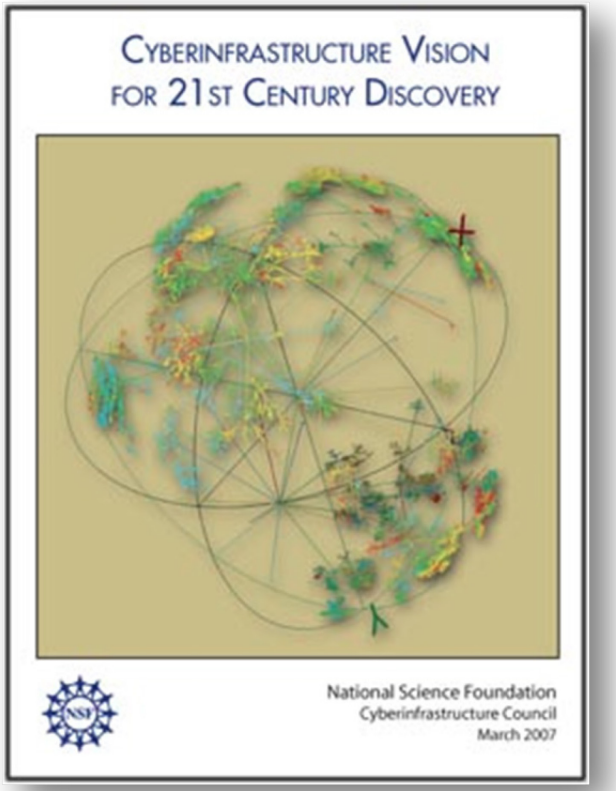
Was this planned, or was it good luck?

NSF sent out a variety of solicitations. Outcome:
3+ winning proposals to different solicitations:

- XSEDE (~2 FTEs devoted to biology through NIP program)
- iPlant (~6 FTEs)
- National Center for Genome Analysis (~4 FTEs)

The winners pieced all of this together. It was partly good luck, partly cross funding, mostly effective collaboration. This was not an outcome the NSF planned in advance. They counted on the community to respond sensibly and effectively.

This time we did



Supporting national research communities: A fundamentally different challenge today

Many more researchers use advanced computing today, and in more disciplines.

There is much less variety in basic underlying processors, but much more variety in packaging (HPC systems to local clusters to clouds)

There is much more political pressure and no general equivalent today for Hilbert's list of mathematical problems.

Current grand challenges	EU	US
Understanding human brain	X (FET Flagship)	X
Big Data	X	X
Global climate change		?
Tokamak design		?
Graphene	X (FET Flagship)	

The RNA transcriptome of the fruit fly was not on anyone's list of grand challenge problems (even though the human genome was!)

Returning to earlier contentions

- Serving national communities is best done by multiple, interacting projects.
- We need proactive and reactive organizations!
- Software innovation and software support are two different processes.
- Things are difficult given politics of science. One best thing we can do with infrastructure is support the scientific community as it is (vs. our ideal of it).
- There is a delicate interaction of funding agencies generating solicitations and centers responding to them.
    - Tensions between supporting “the best science” and “persistence of centers”
        - How do you manage sustainability of groups operating at different layers of infrastructure?
        - If you don’t, how to maintain supply of interested, talented workers?
        - ***Institutionally, it’s critical to continue support for 3 people even though “just” 3 out of 100+***
    - How do you not over-promise? How do you avoid consensus opinions: “O.K., this problem is solved and no longer needs funding.”
- Cloud computing, Internet2 NET+, etc. may change money flow and sustainability strategies

We need “big science,” but must support innovative science even when not big.

Lessons, part 1 – Differences that may be interesting and notable

In the US, competition for funding has, in past times, been a zero-sum game.

Large, collaborative grants are starting to change this in the US.
(Financial limitations in the future will change this even more.)

The greater role of co-funding within the EU might make it easier in the EU than in the US to better plan and coordinate multiple projects.

This makes the game less of a zero sum game and implies more local and regional control.

Lessons, part 2 – Information potentially transferable to EU

We have coordinated excellent support for a given community with some 12 FTEs (Full Time Equivalents) focused on that community. In a national context, it was not that costly to make a big difference to the community served.

Despite some degree of competition, this partnership has become formal and so far successful.

Genome data in general is open, and certain types are well supported and sustained internationally. Indiana University simply plunged ahead and offered data storage persistently, specifically for this community.

It did not take all that much money to make a real difference to the genome research community.

There are no magic bullets. Currently there is no alternative to constantly working for funding.

Thanks!

- The research described here was supported by a number of grant awards:
  - XSEDE: 1053575
  - iPlant: NSF DBI-1265383
  - NCGAS: 1062432
  - Pervasive Technology Institute: Supported by a generous grant from the Lilly Endowment, Inc. and Indiana University

- Any opinions expressed here are those of the speaker and do not necessarily reflect any views held by any of these agencies

- Thanks to the staff of OVPIT and especially PTI and the Research Technologies Division of University Information Technology Services.

- Thanks especially RT Directors / Senior Leaders (Eric Wernert, Matt Link, Therese Miller, Bill Barnett) and Managers (most especially Stephen Simms, Robert Henschel, Richard LeDuc) and NCGAS staff.

Questions and answers from Presentation

Question:

Would it have been possible to get funding for the National Center for Genome Analysis Support at the State level, rather than at the national level?

Answer:

With the US funding models, it would not have been possible to get funding at less than the national level. There are no programs for funding an activity like this at the State level in the US. However, if it were possible to get funding, for example, at the level of one of the German Bundeslaender, a very small group of people – say just two perhaps – could make a real difference in research within that Land.

License Terms

- Please cite as: Stewart, C.A. 2014. Serving national scientific communities - genome analysis as an example. ZKI-Frühjahrstagung, Berlin, Deutschland. 25 March 2014. http://http://hdl.handle.net/2022/17383
- Items indicated with a © are under copyright and used here with permission. Such items may not be reused without permission from the holder of copyright except where license terms noted on a slide permit reuse.
- Except where otherwise noted, contents of this presentation are copyright 2013 by the Trustees of Indiana University.
- This document is released under the Creative Commons Attribution 3.0 Unported license (http://creativecommons.org/licenses/by/3.0/). This license includes the following terms: You are free to share – to copy, distribute and transmit the work and to remix – to adapt the work under the following conditions: attribution – you must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work). For any reuse or distribution, you must make clear to others the license terms of this work.